GHENT UNIVERSITY

imec

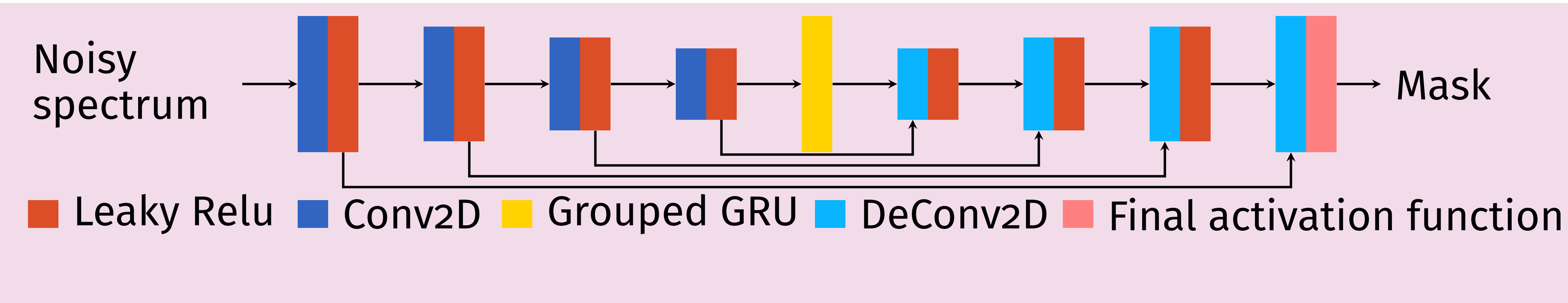
PHASE RECONSTRUCTION IN SINGLE CHANNEL SPEECH ENHANCEMENT BASED ON PHASE GRADIENTS AND ESTIMATED CLEAN-SPEECH AMPLITUDES

Yanjue Song, Nilesh Madhu

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium

{yanjue.song, nilesh.madhu}@ugent.be

1 Introduction



$$S(l,m) = |S(l,m)| \exp\left(j\Phi_S(l,m)\right)$$

- Phase estimation in single-channel speech enhancement
  - Taking the noisy phase: the MMSE-sense optimal
  - Complex mask/mapping
- **Challenge**: no pattern observed
- Pattern exists in phase gradients

$$\Delta_f\Phi(l,m) = \Phi(l,m) - \Phi(l,m-1) \tag{1}$$
$$\Delta_t\Phi(l,m) = \Phi(l,m) - \Phi(l-1,m) \tag{2}$$

→ Estimate phase from clean magnitude (phase retrieval)

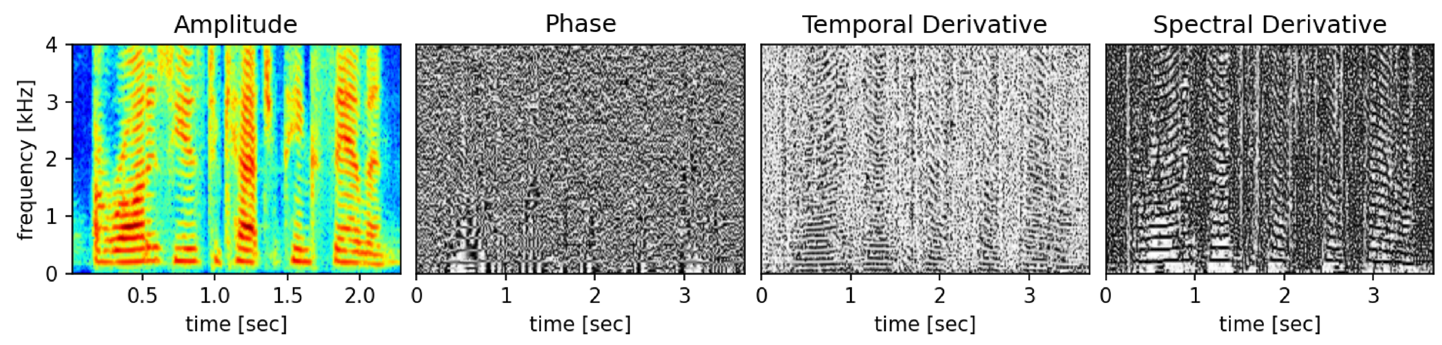


Figure 1: The amplitude, phase, and phase derivative of the clean speech.

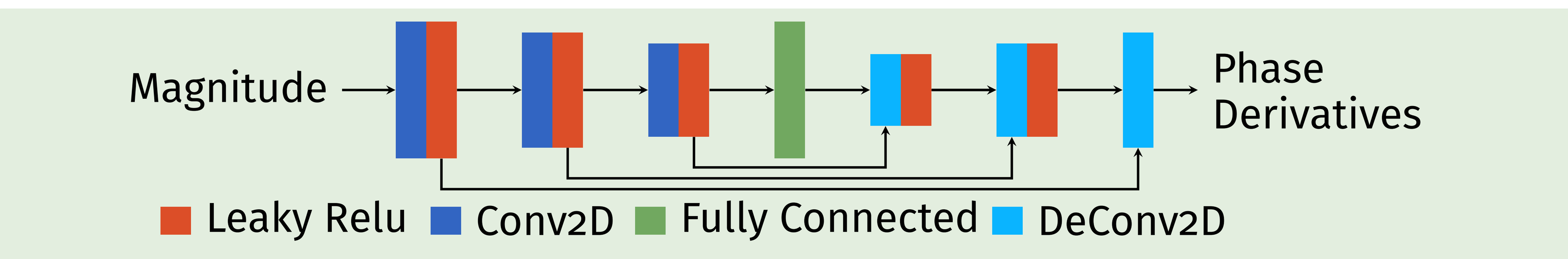
2 Motivation

- Can we utilise the gradient information in speech enhancement?
- Phase retrieval solution: too artificial
- What's missing: the initial phase estimate
- **Idea**: ground it by the initial phase estimate
- Our method: fuse

(a) Phase estimate from $\Delta_t\Phi(l,m)$ (*temporal* derivative)
(b) Phase estimate from $\Delta_f\Phi(l,m)$ (*spectral* derivative)
(c) **Initial phase estimate**

to get *one consistent phase estimate* for the enhanced speech

3 Phase Derivative Estimation



- Loss function: $\mathcal{L}_\star = \sum_{l,m}\left(1 - \cos\left(\widehat{\Delta_\star\Phi(l,m)} - \Delta_\star\Phi(l,m)\right)\right), \star \in \{t,f\}$
- Training scheme: **matched** or **agnostic**?

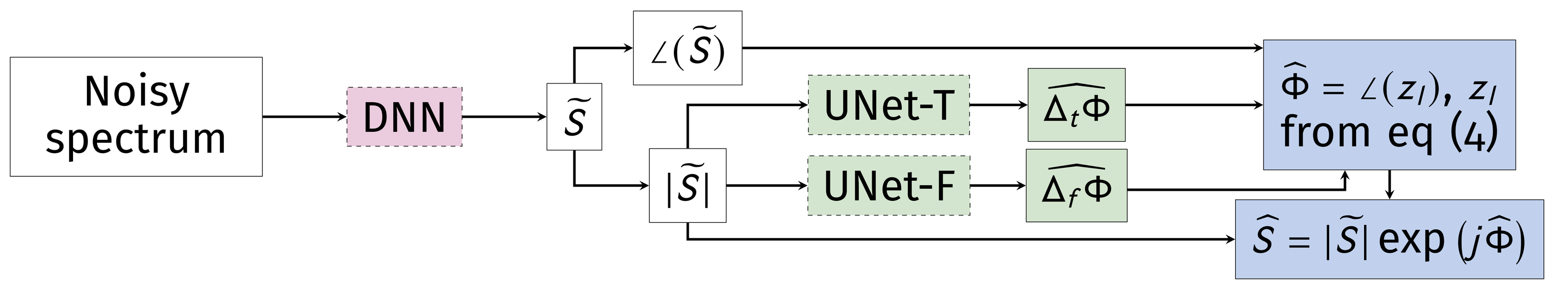
4 Phase Reconstruction

- Goal: one consistent phase estimate based on the three different estimates
- Minimise the distance between this final estimate z_l and all sources
- Translated to the cost function:

$$\mathcal{J}(z_l) = \underbrace{\left\|z_l - \widehat{V}_l \odot \widehat{S}_{l-1}\right\|^2_{\Lambda_l}}_{\text{distance to } \widehat{\Phi} \text{ from } \widehat{\Delta_t\Phi}} + \underbrace{\left\|D_l z_l\right\|^2_{\Gamma_l}}_{\text{distance to } \widehat{\Phi} \text{ from } \widehat{\Delta_f\Phi}} + \underbrace{\left\|z_l - \widetilde{S}_l\right\|^2_{\Omega_l}}_{\text{distance to } \widetilde{S}}, \tag{3}$$

where z is the clean speech estimate.
→ The optimal solution:

$$\widehat{z}_l = \left(\Lambda_l + D_l^H\Gamma_l D_l + \Omega_l\right)^{-1}\left(\Lambda_l(\widehat{V}_l \odot \widehat{S}_{l-1}) + \Omega_l\widetilde{S}_l\right) \tag{4}$$



5 Evaluation Results

Dataset

- Training: DNS challenge 2021, 140 hours
- Test: DNS challenge 2020, synthetic test set

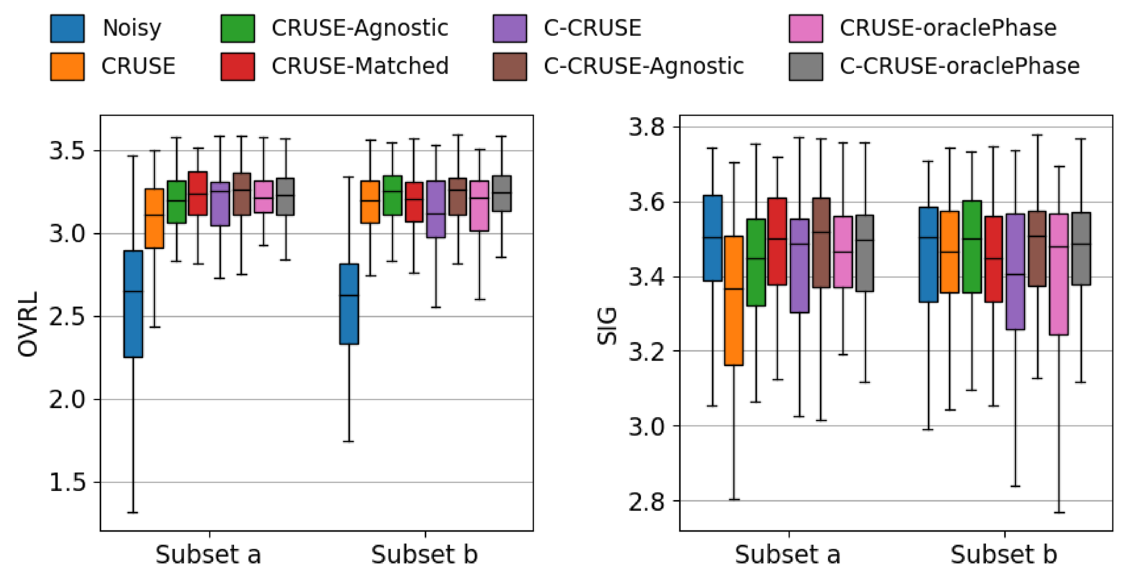
Objective metrics

Table 1: Averaged instrumental metrics on test set. Best results in bold.

Method	segSNR [dB]	STOI	DNSMOS	
			OVRL	SIG
Noisy	6.87	0.87	2.53	3.33
CRUSE	13.74	0.93	3.10	3.36
CRUSE-Agnostic	14.30	0.93	3.17	3.43
CRUSE-Matched	14.19	0.93	3.17	3.44
C-CRUSE	13.92	0.93	3.14	3.40
C-CRUSE-Agnostic	**14.45**	0.93	**3.20**	**3.45**
CRUSE-OraclePhase	14.51	0.94	3.17	3.43
C-CRUSE-OraclePhase	14.77	0.94	3.20	3.45

DNSMOS Distribution

- subset **a)** mixtures with *stationary* or *short-term stationary* noise
- subset **b)** mixtures with *sparse, transient* noise



- Phase enhancement
  - improves all metrics
  - is comparable to using *oracle* phase
- Boosts signal quality in poor SNR conditions
- For stationary noise: SE-Matched > SE-Agnostic
- For sparse noise: SE-Agnostic > SE-Matched

Spectrogram Samples

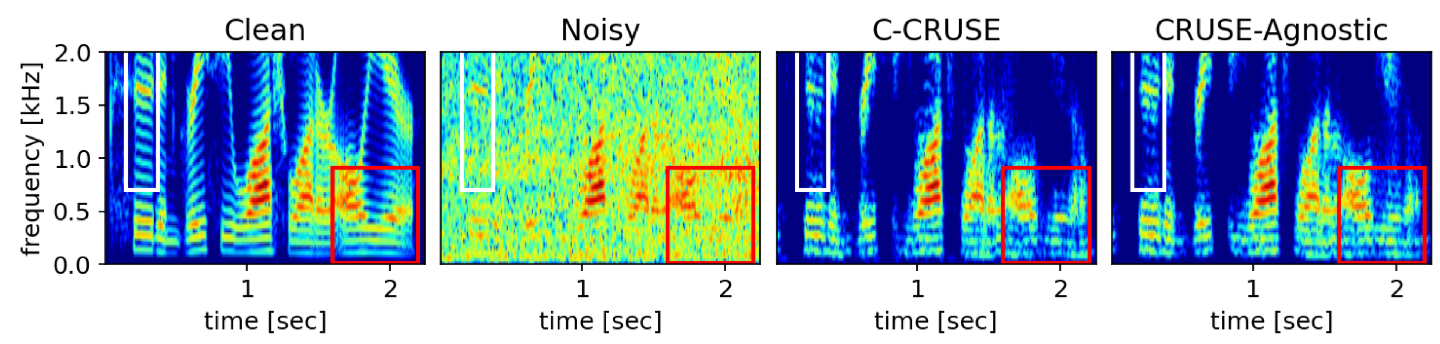


Figure 2: Noisy signal: Street noise, -2 dB.

- Similar performance in high SNR regions
- More continuous harmonics by the proposed phase reconstruction

6 Conclusions

- Incorporating initial phase estimate for natural-sounding output
- Improvement reflected in objective audio quality metrics
- Compatible with real- or complex-domain methods
- Matched/agnostic methods suit different noise types